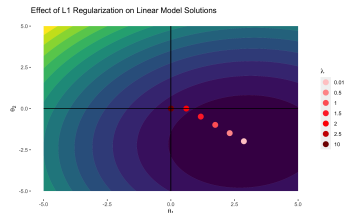


# Introduction to Machine Learning

## Regularization

## Lasso Regression



### Learning goals

- Lasso regression /  $L_1$  penalty
- Know that lasso selects features
- Support recovery

# LASSO REGRESSION

Another shrinkage method is the so-called **lasso regression** (least absolute shrinkage and selection operator), which uses an  $L_1$  penalty on  $\theta$ :

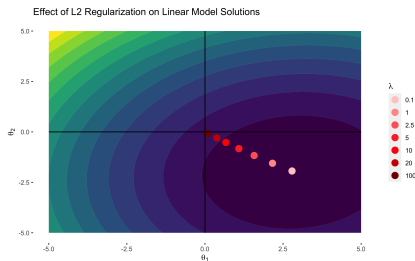
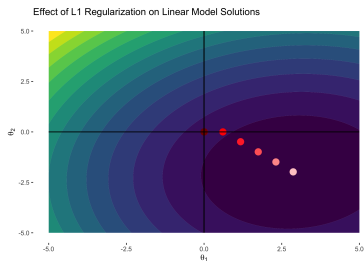
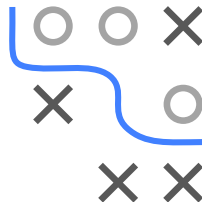
$$\begin{aligned}\hat{\theta}_{\text{lasso}} &= \arg \min_{\theta} \sum_{i=1}^n \left( y^{(i)} - \theta^T \mathbf{x}^{(i)} \right)^2 + \lambda \sum_{j=1}^p |\theta_j| \\ &= \arg \min_{\theta} (\mathbf{y} - \mathbf{X}\theta)^T (\mathbf{y} - \mathbf{X}\theta) + \lambda \|\theta\|_1\end{aligned}$$

Optimization is much harder now.  $\mathcal{R}_{\text{reg}}(\theta)$  is still convex, but in general there is no analytical solution and it is non-differentiable.



# LASSO REGRESSION / 2

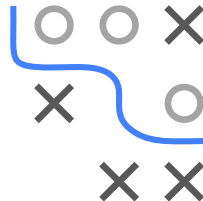
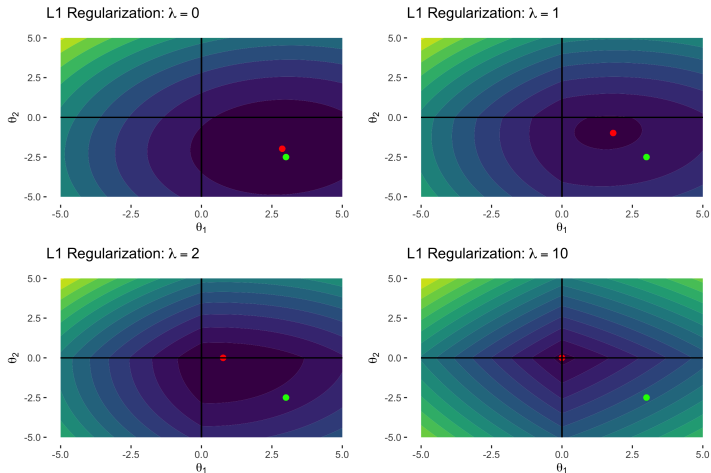
Let  $y = 3x_1 - 2x_2 + \epsilon$ ,  $\epsilon \sim N(0, 1)$ . The true minimizer is  $\theta^* = (3, -2)^T$ . LHS =  $L1$  regularization; RHS =  $L2$



With increasing regularization,  $\hat{\theta}_{lasso}$  is pulled back to the origin, but takes a different “route”.  $\theta_2$  eventually becomes 0!

# LASSO REGRESSION / 3

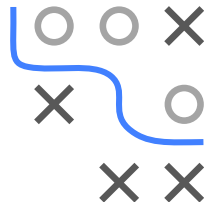
Contours of regularized objective for different  $\lambda$  values.



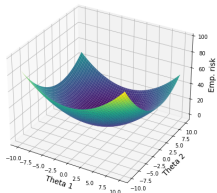
Green = true minimizer of the unreg. objective and red = lasso solution.

# LASSO REGRESSION / 4

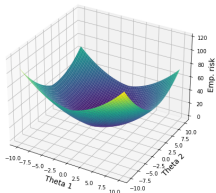
Regularized empirical risk  $\mathcal{R}_{\text{reg}}(\theta_1, \theta_2)$  using squared loss for  $\lambda \uparrow$ .  $L1$  penalty makes non-smooth kinks at coordinate axes more pronounced, while  $L2$  penalty warps  $\mathcal{R}_{\text{reg}}$  toward a “basin” (elliptic paraboloid).



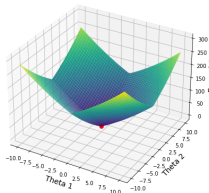
Regularization: L1, Lambda: 0



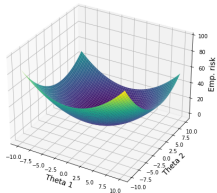
Regularization: L1, Lambda: 1



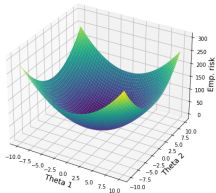
Regularization: L1, Lambda: 10



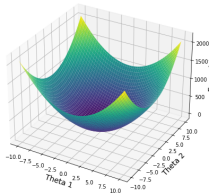
Regularization: L2, Lambda: 0



Regularization: L2, Lambda: 1



Regularization: L2, Lambda: 10

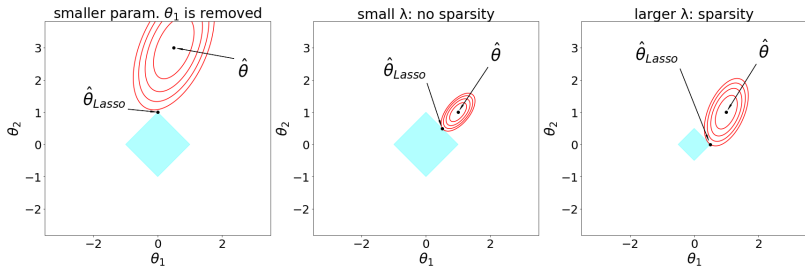
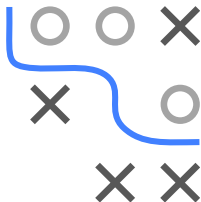


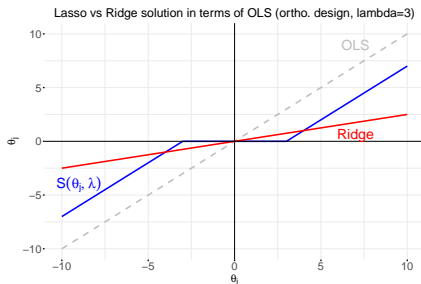
# LASSO REGRESSION / 5

We can also rewrite this as a constrained optimization problem. The penalty results in the constrained region to look like a diamond shape.

$$\min_{\theta} \sum_{i=1}^n \left( y^{(i)} - f(\mathbf{x}^{(i)} | \theta) \right)^2 \text{ subject to: } \|\theta\|_1 \leq t$$

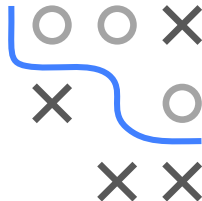
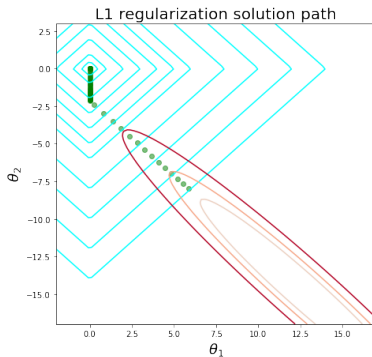
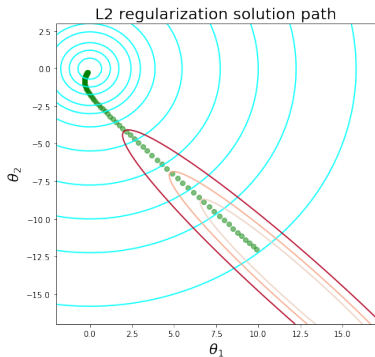
The kinks in  $L1$  enforce sparse solutions because “the loss contours first hit the sharp corners of the constraint” at coordinate axes where (some) entries are zero.



$$\hat{\theta}_{\text{lasso}} = \text{sign}(\hat{\theta}_{\text{OLS}})(|\hat{\theta}_{\text{OLS}}| - \lambda)_+ \quad (\text{sparsity})$$
$$\hat{\theta}_{\text{Ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y} = ((1 + \lambda) \mathbf{I})^{-1} \hat{\theta}_{\text{OLS}} = \frac{\hat{\theta}_{\text{OLS}}}{1 + \lambda} \quad (\text{no sparsity})$$


# COMPARING SOLUTION PATHS FOR $L1/L2$

- Ridge results in smooth solution path with non-sparse params
- Lasso induces sparsity, but only for large enough  $\lambda$





# SUPPORT RECOVERY OF LASSO

► Zhao and Yu 2006

When can lasso select true support of  $\theta$ , i.e., only the non-zero parameters?

Can be formalized as sign-consistency:

$$\mathbb{P}(\text{sign}(\hat{\theta}) = \text{sign}(\theta)) \rightarrow 1 \text{ as } n \rightarrow \infty \quad (\text{where } \text{sign}(0) := 0)$$

Suppose the true DGP given a partition into subvectors  $\theta = (\theta_1, \theta_2)$  is

$$Y = X\theta + \varepsilon = X_1\theta_1 + X_2\theta_2 + \varepsilon \text{ with } \varepsilon \sim (0, \sigma^2 I)$$

and only  $\theta_1$  is non-zero. Let  $X_1$  denote the  $n \times q$  matrix with the relevant features and  $X_2$  the matrix of noise features. It can be shown that  $\hat{\theta}_{\text{lasso}}$  is sign consistent under an **irrepresentable condition**:

$$|(\mathbf{X}_2^\top \mathbf{X}_1)(\mathbf{X}_1^\top \mathbf{X}_1)^{-1} \text{sign}(\theta_1)| < \mathbf{1} \text{ (element-wise)}$$

In fact, lasso can only be sign-consistent if this condition holds.

Intuitively, the irrelevant variables in  $X_2$  must not be too correlated with (or *representable* by) the informative features

► Meinshausen and Yu 2009

